Incorporating W3C's DQV and PROV in CISER's Data Quality Review and Reproduction of Results Service

William C. Block Florio O. Arguillas Jeremy Williams
Cornell Institute for Social and Economic Research (CISER)

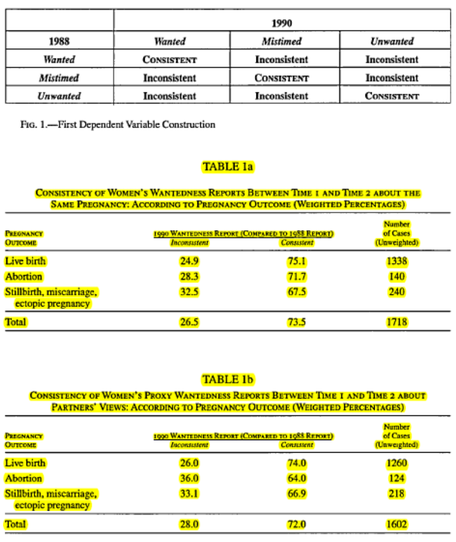
Corresponding author e-mail:
block@cornell.edu

Abstract

A year and a half after its implementation, CISER's Research Data Quality Review and Reproduction of Results Service (or R^2, for short), a service developed to encourage sharing of high quality data, code, documentation, and metadata associated with a study for the purpose of reproducible research, continues to evolve and improve. This poster discusses: a) the service at its current state; b) the improvements made to the service including cost-reduction and buy-in strategies to encourage researchers to use the service; c) pre- and post-reproduction services to improve data, code, documentation, and metadata quality; d) the inclusion of the Cornell's CED^2AR software for assessing and generating complete DDI metadata; and e) the utilization of the CISER Data Archive as the free and permanent home for the study and its associated files. Last, our poster will include for the first time our efforts to incorporate the W3C's Data Quality Vocabulary (DQV) and PROV metadata into our R^2 service. DQV provides a means to describe the subjective measures applied to ensure the integrity of data sources and introduces a way of expressing guidelines for data quality that producers of data can abide by. The W3C PROV ontology defines entities, agents, and activities related to the origin of resources, which is particularly important in the common scenario of referencing multiple datasets for a given investigation." By incorporating DQV and PROV into the R^2 data management workflow, we hope to provide a foundation of support improving the quality of research inputs and outputs, which in turn can have a chain effect for derivative data products.

Golden Rule of CISER's Replication Service:

Output produced by running code against the data should be identical to the publication up to the last decimal place. Slight deviation is not acceptable and must be investigated.



	1990		
1988	*Wanted*	*Mistimed*	*Unwanted*
Wanted	CONSISTENT	Inconsistent	Inconsistent
Mistimed	Inconsistent	CONSISTENT	Inconsistent
Unwanted	Inconsistent	Inconsistent	CONSISTENT

FIG. 1.—First Dependent Variable Construction

TABLE 1a

CONSISTENCY OF WOMEN'S WANTEDNESS REPORTS BETWEEN TIME 1 AND TIME 2 ABOUT THE SAME PREGNANCY: ACCORDING TO PREGNANCY OUTCOME (WEIGHTED PERCENTAGES)

Pregnancy Outcome	1990 Wantedness Report (Compared to 1988 Report) Inconsistent	Consistent	Number of Cases (Unweighted)
Live birth	24.9	75.1	1338
Abortion	28.3	71.7	140
Stillbirth, miscarriage, ectopic pregnancy	32.5	67.5	240
Total	26.5	73.5	1718

TABLE 1b

CONSISTENCY OF WOMEN'S PROXY WANTEDNESS REPORTS BETWEEN TIME 1 AND TIME 2 ABOUT PARTNERS' VIEWS: ACCORDING TO PREGNANCY OUTCOME (WEIGHTED PERCENTAGES)

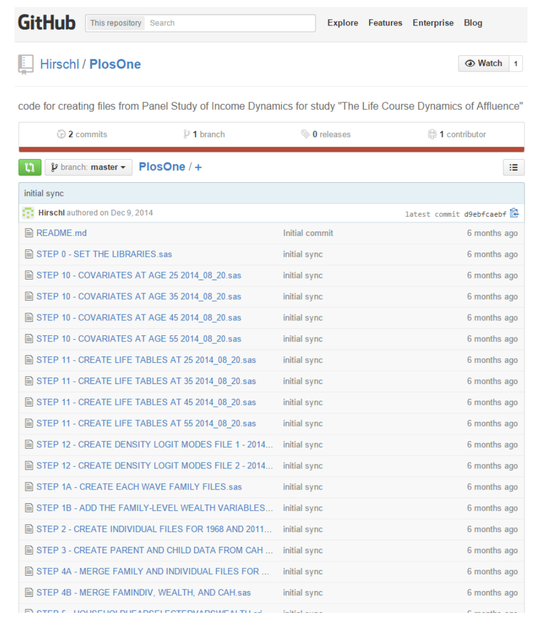
Pregnancy Outcome	1990 Wantedness Report (Compared to 1988 Report) Inconsistent	Consistent	Number of Cases (Unweighted)
Live birth	26.0	74.0	1260
Abortion	36.0	64.0	124
Stillbirth, miscarriage, ectopic pregnancy	33.1	66.9	218
Total	28.0	72.0	1602

rately for different Time 1 intention statuses. As a result, in those two tables the dependent variable can capture shifts toward more positive and more negative views. The dependent variable comprises three categories: More positive, more negative, and consistent. A more positive report resulted (1) if the woman claimed at Time 1 that the pregnancy was unwanted, and then reported at Time 2 that the pregnancy was mistimed or intended, or (2) if she said at Time 1 that the pregnancy was mistimed and at Time 2 that it was intended. A more negative report resulted (1) if the respondent said at Time 1 that the pregnancy was intended and said

```
/* Survey set the data to incorporate PSUs and strata*/
svyset [pweight=perswgt], strata(sest) psu(secu)

/* Table 1 descriptives-Overall sample*/
/* 1995 vs. 2006-10 women descriptives (col's 1 and 2 in T1, descriptives)*/
bys survey:ci duration ucon2 ucon3 cohabbirth black hispan other dadedu3 dadedu2 dadedu1  dadedu4 ///
momedu3 momedu2 momedu1 momedu4 intact church ///
lths high somecoll coll agemomy agedady planned momprevkids priormar priorcoh anyprioruni  ///
 mistimed unwanted ageoldest1 ageoldest2 ageoldest3 babdoby prioruni kids timebaby  anyshotgun ///
if b10==1 [aweight=perswgt]

/*Table 2: Logistic regressions*/
/*Model 1*/
local alttvmar "newtvmar"
foreach j in 1995 2007{
svy: logit tvsplit tvduration cohabunion cohabbirth `alttvmar' ///
if b10==1 & survey=="j' ,or

/*Test coefficients in Table 3*/
/* Row 1*/
/*CMB vs MB*/
test [t1995_tvsplit]cohabunion+[t1995_tvsplit]`alttvmar'=[t2007_tvsplit]cohabunion+[t2007_tvsplit]`altt
test [t1995_tvsplit]cohabunion=0
test [t2007_tvsplit]cohabunion=0

/*C-B-M vs MB*/
test [t2007_tvsplit]cohabunion+[t2007_tvsplit]cohabbirth+[t2007_tvsplit]`alttvmar'=[t1995_tvsplit]cohab
test 0=[t1995_tvsplit]cohabunion+[t1995_tvsplit]cohabbirth
test 0=[t2007_tvsplit]cohabunion+[t2007_tvsplit]cohabbirth

/*CB vs MB*/
test [t1995_tvsplit]cohabunion+[t1995_tvsplit]cohabbirth=[t2007_tvsplit]cohabunion+[t2007_tvsplit]cohab
test [t1995_tvsplit]`alttvmar'=[t1995_tvsplit]cohabunion+[t1995_tvsplit]cohabbirth
test [t2007_tvsplit]`alttvmar'=[t2007_tvsplit]cohabunion+[t2007_tvsplit]cohabbirth

/* Row 2 1995 vs Row 2 2006-10*/
test - [t1995_tvsplit]cohabbirth =- [t2007_tvsplit]cohabbirth
test - [t1995_tvsplit]cohabbirth +[t1995_tvsplit]`alttvmar' = - [t2007_tvsplit]cohabbirth +[t2007_tvs
```

Common problems:

- Very long, complex codes
- Unnecessary/excess sections of codes whose outputs are not found in the paper (this delays replication because the Staff has to go through the entire code and its output, and figure out where they are on the paper)
- Code points to subdirectories for retrieving or saving data, thus Staff has to recreate the directory structure for the code to run correctly
- Some codes are not efficient, but <u>will not</u> be modified by the Staff. The Staff, however, will suggest ways to make it efficient.
- Codes are often multiple files with no indication of sequence. Reproducer has to determine which to run first especially if codes build on top of the other.



GitHub This repository Search Explore Features Enterprise Blog

Hirschl / PlosOne Watch 1

code for creating files from Panel Study of Income Dynamics for study "The Life Course Dynamics of Affluence"

2 commits 1 branch 0 releases 1 contributor

branch: master PlosOne / +

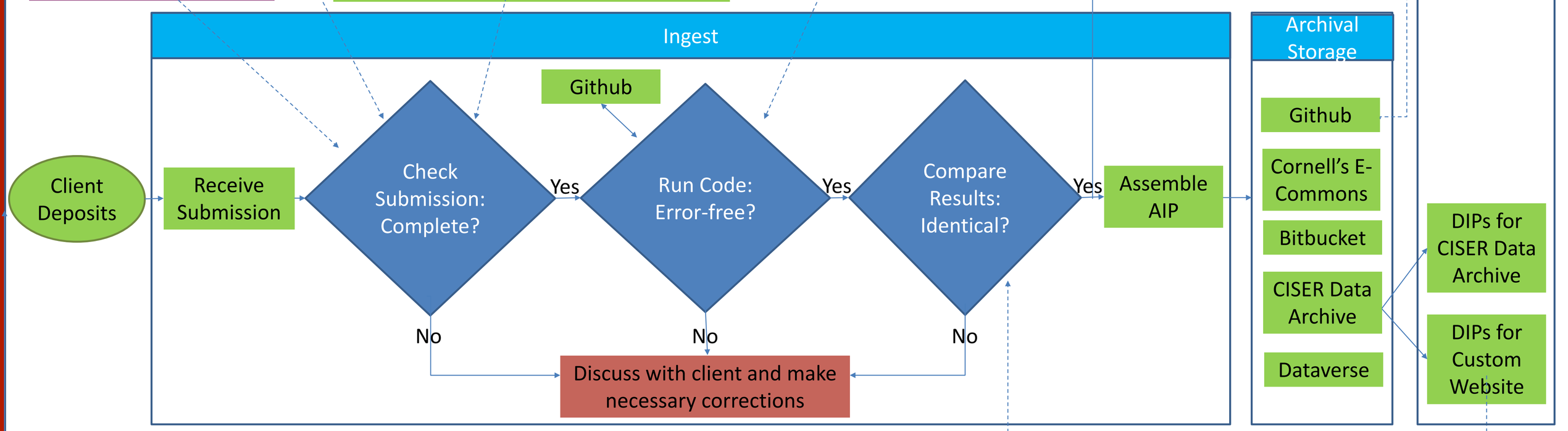
initial sync

Hirschl authored on Dec 9, 2014 latest commit d9ebfcaebf

README.md	Initial commit	6 months ago
STEP 0 - SET THE LIBRARIES.sas	initial sync	6 months ago
STEP 10 - COVARIATES AT AGE 25 2014_08_20.sas	initial sync	6 months ago
STEP 10 - COVARIATES AT AGE 35 2014_08_20.sas	initial sync	6 months ago
STEP 10 - COVARIATES AT AGE 45 2014_08_20.sas	initial sync	6 months ago
STEP 10 - COVARIATES AT AGE 55 2014_08_20.sas	initial sync	6 months ago
STEP 11 - CREATE LIFE TABLES AT 25 2014_08_20.sas	initial sync	6 months ago
STEP 11 - CREATE LIFE TABLES AT 35 2014_08_20.sas	initial sync	6 months ago
STEP 11 - CREATE LIFE TABLES AT 45 2014_08_20.sas	initial sync	6 months ago
STEP 11 - CREATE LIFE TABLES AT 55 2014_08_20.sas	initial sync	6 months ago
STEP 12 - CREATE DENSITY LOGIT MODES FILE 1 - 2014...	initial sync	6 months ago
STEP 12 - CREATE DENSITY LOGIT MODES FILE 2 - 2014...	initial sync	6 months ago
STEP 1A - CREATE EACH WAVE FAMILY FILES.sas	initial sync	6 months ago
STEP 1B - ADD THE FAMILY-LEVEL WEALTH VARIABLES...	initial sync	6 months ago
STEP 2 - CREATE INDIVIDUAL FILES FOR 1968 AND 2011...	initial sync	6 months ago
STEP 3 - CREATE PARENT AND CHILD DATA FROM CAH ...	initial sync	6 months ago
STEP 4A - MERGE FAMILY AND INDIVIDUAL FILES FOR ...	initial sync	6 months ago
STEP 4B - MERGE FAMINDIV, WEALTH, AND CAH.sas	initial sync	6 months ago

Researcher provides CISER Staff copy of article, code(s), and data; highlights the sections on the article with figures derived from running code against data; and put comments on the codes that describe what the section of the code will produce or is doing. CISER uses CED2AR to check for completeness and create DDI metadata

Common problems: No variable and value labels



Data Management

Build Metadata (DDI, DQV, PROV)

Metadata Database

Access

CED2AR

Ingest

Client Deposits

Receive Submission

Check Submission: Complete?

Yes

Run Code: Error-free?

Yes

Compare Results: Identical?

Yes

Assemble AIP

Github

No

No

No

Discuss with client and make necessary corrections

Archival Storage

Github

Cornell's E-Commons

Bitbucket

CISER Data Archive

Dataverse

DIPs for CISER Data Archive

DIPs for Custom Website

Cost-reduction strategies

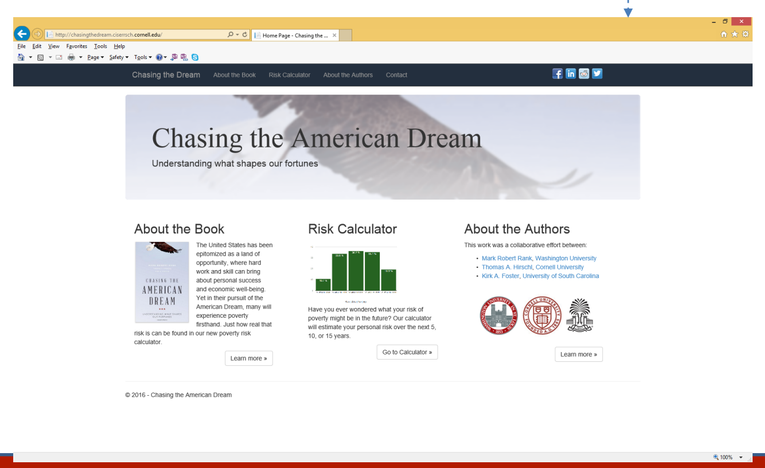
- Data curation and management training
- Code writing and organization training
- Code efficiency training e.g., macro programming, SQL programming
- Version control software training e.g., Github

iassist

International Association for Social Science Information Services & Technology

Common problems:

- Some results do not match article
- Not all figures printed on the article are produced by the code. Some involved other software packages such as Excel.
- Order of variables in the model in the printed table do not match the order of the variables in the output table for that model, which slows down the verification process.



Chasing the Dream About the Book Risk Calculator About the Authors Contact

Chasing the American Dream

Understanding what shapes our fortunes

About the Book

The United States has been epitomized as a land of opportunity, where hard work and skill can bring about personal success and economic well-being. Yet in their pursuit of the American Dream, many will experience poverty firsthand. Just how real that risk is can be found in our new poverty risk calculator.

Learn more »

Risk Calculator

Have you ever wondered what your risk of poverty might be in the future? Our calculator will estimate your personal risk over the next 5, 10, or 15 years.

Go to Calculator »

About the Authors

This work was a collaborative effort between:

- Mark Robert Rank, Washington University
- Thomas A. Hirschl, Cornell University
- Kirk A. Foster, University of South Carolina

Learn more »

© 2016 - Chasing the American Dream